HIPERGRÁFOK EUKLIDESZI TÉRBE VALÓ BEÁGYAZÁSA VELESZÜLETETT RENDELLENESSÉGEK CLUSTEREZÉSÉHEZ

Bolla Marianna, Tusnády Gábor

MTA Számitástechnikai és Automatizálási Kutató Intézet, MTA Matematikai Kutató Intézet

A modell

Legyen H egy hipergráf $P_1,\dots,P_n$ pontokkal és $G_1,\dots,G_m$ élekkel, ahol az élek a pontok bizonyos részhalmazai. A hipergráfot megadhatjuk egy nxm-es A indicenciamátrix-szal, melyre

$$a_{ji} = \begin{cases} 1, \text{ ha } P_j \in G_i, \\ 0, \text{ ha } P_j \notin G_i, \end{cases} \qquad j=1,..,n;\ i=1,...m\ .$$

Keresünk olyan $\underline{x}_1,\dots,\underline{x}_n \in E_k$ és $\underline{y}_1,\dots,\underline{y}_m \in E_k$ pontokat /ezek lesznek a H hipergráf pontjainak ill. éleinek képei a k-dimenziós valós euklideszi térben , melyekre

$$Q = \sum_{j=1}^{n} \sum_{i=1}^{m} a_{ji} \| \underline{x}_j - \underline{y}_i \|^2 \qquad (1)$$

minimális, ahol $k < \min\{n,m\}$ előre adott természetes szám. Azaz minden élre vesszük az él képének a tartalmazott pontok képeitől való távolságának négyzetét, és ezt összegezzük az összes élre. Megmutatjuk, hogy a fenti célfüggvény minimalizálására bizonyos mellékfeltételekkel explicit megoldás adható. Célunk k-t olyan "kicsinek" választani, hogy az $\underline{x}_1,\dots,\underline{x}_n$ ill. $\underline{y}_1,\dots,\underline{y}_m$ pontok a k-dimenziós euklideszi térben jól clusterezhetők legyenek /pl. az ottani távolságok alapján/.

A feladat megoldása

Vezessük be a következő jelöléseket:

$$X = (\underline{x}_1,\dots,\underline{x}_n)\ , \quad Y = (\underline{y}_1,\dots,\underline{y}_m)$$

a képvektorokat oszloponként tartalmazó kxn-es ill. kxm-es mátrixok.

$$S = \begin{bmatrix} s_1 & & 0 \\ & s_2 & \\ & & \ddots \\ 0 & & s_n \end{bmatrix}, \text{ ahol } s_j = \sum_{i=1}^{m} a_{ji}, \ (j=1,\dots,n)$$

$$T=\begin{bmatrix} t_1 & & 0 \\ & t_2 & \\ & & \ddots \\ 0 & & t_m \end{bmatrix}, \text{ ahol } t_i = \sum_{j=1}^{n} a_{ji}, \ (i=1,\ldots,m).$$

Tegyük fel, hogy $t_i > 0$, $(i=1,\ldots,m)$, azaz minden él tartalmaz pontot. A pontok képeire tegyük továbbá a következő kényszerfeltételt:

$$XX^T = \sum_{j=1}^{n} \underline{x}_j \underline{x}_j^T = I_k \qquad (2)$$

(vagyis X sorai ortonormált rendszert alkotnak.)
A megoldáshoz alakitsuk át a Q kifejezést:

$$Q= \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ji} \| \underline{x}_j \|^2 + \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ji} \| \underline{y}_i \|^2 - 2 \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ji} \underline{y}_i^T \underline{x}_j =$$

$$= \sum_{j=1}^{n} s_j \| \underline{x}_j \|^2 + \sum_{i=1}^{m} t_i \| \underline{y}_i \|^2 - 2 \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ji} \underline{y}_i^T \underline{x}_j .$$

Először rögzitsük az $\underline{x}_j$ $(j=1,\ldots,n)$ vektorokat és fejezzük ki az $\underline{y}_i$ $(i=1,\ldots,m)$ vektorokat ezek segitségével! Ehhez deriváljuk a fenti kifejezést $\underline{y}_i$ szerint. A derivált

$$2t_i\underline{y}_i - 2 \sum_{j=1}^{n} a_{ji}\underline{x}_j$$

lesz, amit zérussá téve

$$\underline{y}_i = \frac{1}{t_i} \sum_{j=1}^{n} a_{ji}\underline{x}_j, \ (i=1,\ldots,m) \qquad (3)$$

adódik, azaz egy él képe az általa tartalmazott pontok képének a sulypontja. Ezt visszairva Q kifejezésébe:

$$Q= \sum_{j=1}^{n} s_j \| \underline{x}_j \|^2 + \sum_{i=1}^{m} t_i \| \underline{y}_i \|^2 - 2 \sum_{i=1}^{m} \underline{y}_i^T t_i \underline{y}_i$$

$$= \sum_{j=1}^{n} \underline{s}_j \| \underline{x}_j \|^2 - \sum_{i=1}^{m} t_j \| \underline{y}_i \|^2 \qquad (4)$$

$$= \mathrm{tr}\,(XSX^T - YTY^T) = \mathrm{tr}\{X(S-AT^{-1}A^T)\,X^T\},$$

ahol kihasználtuk, hogy $Y=XAT^{-1}$ a (3) összefüggés miatt és T invertálható.

Tehát a (4) kifejezést kell X oszlopai szerint minimalizálni a (2) mellékfeltétellel. Ebből Y oszlopai már (3) szerint adódnak. Mivel X^T oszlopai ortonormált rendszert alkotnak és az $S-AT^{-1}A^T$ mátrix

szimmetrikus, pozitiv szemidefinit, a feladat megoldását a következő lemma alapján kapjuk:

Lemma: Ha B nxn-es szimmetrikus, pozitiv szemidefinit adott mátrix, Z pedig olyan nxk-as mátrix (k≤n adott egész), melynek oszlopai E_n-ben ortonormált rendszert alkotnak, akkor

$$\sum_{i=n-k+1}^{n} \lambda_i \leq \mathrm{tr}(Z^T B Z) \leq \sum_{i=1}^{k} \lambda_i ,$$

és a maximum /minimum/ olyan Z mátrixra vétetik fel, mely a B mátrix k legnagyobb /legkisebb/ sajátértékéhez tartozó normált sajátvektort tartalmazza oszlopaiban, ahol $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ a B mátrix sajátértékei.

Igy $B=S-AT^{-1}A^T$, $Z=X^T$ jelöléssel az (1) kifejezés minimuma

$$\sum_{i=n-k+1}^{n} \lambda_i ,$$

ahol $\lambda_1 \geq \ldots \geq \lambda_n$ a B mátrix sajátértékei. Ha $\underline{u}_1,\ldots,\underline{u}_n$ jelöli a hozzájuk tartozó normált sajátvektorokat, akkor a megoldást adó X mátrix sorait az $\underline{u}_{n-k+1},\ldots,\underline{u}_n$ vektorok alkotják.

Megjegyezzük, hogy a B mátrix legkisebb sajátértéke 0 az $\left(\frac{1}{n},\ldots,\frac{1}{n}\right)^T$ normált sajátvektorral.

Ui: $$\sum_{i=1}^{n} b_{ji} = s_j - \sum_{i=1}^{n}\sum_{\ell=1}^{m} \frac{a_{j\ell}a_{i\ell}}{t_\ell} = \sum_{i=1}^{m} a_{ji} - \sum_{\ell=1}^{m} \frac{a_{j\ell}}{t_\ell} \underbrace{\sum_{i=1}^{n} a_{i\ell}}_{t_\ell} = 0 ,$$

azaz B sorainak összege 0. Igy

$$B\left(\frac{1}{n},\ldots,\frac{1}{n}\right)^T = 0,$$

amiből az állitás következik.

Mivel az utolsó sajátvektor az $(1,\ldots,1)^T$ vektor skalárszorosa, ez azt jelenti, hogy az $\underline{x}_1,\ldots,\underline{x}_n$ képvektorok utolsó koordinátája egyenlő. Igy k=1 esetén az $\underline{x}_1,\ldots,\underline{x}_n$ pontok egybeesnek, aminek alapján, bár Q=0, nem clusterezhetők. Általában is, k>1 esetén a pontok k-1 dimenzióban clusterezhetők jól, igy sikbeli clusterezéshez k=3 kezdőértéket kell választani. Más szóval, ha valódi k-dimenziós clustereket akarunk, akkor X sorait az $\underline{u}_{n-k},\ldots,\underline{u}_{n-1}$ vektorok adják, a minimum értéke pedig $\sum_{i=n-k}^{n-1} \lambda_i$ lesz.

Megjegyezzük, hogy a (2) mellékfeltétel helyett az

$$X \Delta X^T = I_k \qquad (5)$$

mellékfeltétel is használható, ahol Δ tetszőleges nxn-es pozitiv definit mátrix. Ezzel a (4) kifejezés

$$tr\{(X\Delta^{1/2})[\Delta^{-1/2}(S-AT^{-1}A^T)\Delta^{-1/2}](X\Delta^{1/2})^T$$

lesz. Itt az (5) mellékfeltétel miatt $(X\,\Delta^{1/2})^T$ oszlopai alkotnak ortonormált rendszert. A feladat megoldását most is a lemma alapján kapjuk, csak itt a $\Delta^{-1/2}(S-AT^{-1}A^T)\Delta^{-1/2}$ mátrix spektrálfelbontását kell elvégezni, a sajátvektorokat pedig a végén $\Delta^{-1/2}$-nel transzformálni.

A modell alkalmazása

A fenti modellt többszörös veleszületett rendellenességek vizsgálatára használtuk, ahol a hipergráf élei az ujszülöttek, pontjai pedig a rajtuk születéskor megfigyelt rendellenességek voltak. Esetünkben n=40, m=1186742 volt, de látni fogjuk, hogy Q értékét csak a legalább két rendellenességgel rendelkező gyerekek befolyásolják, akik száma 2762. Célunk a rendellenességek, és ezzel párhuzamosan a gyerekek clusterezése volt.

Gráfunk incidenciamátrixa:

$$a_{ji} = \begin{cases} 1, & \text{ha az i-edik gyereken megfigyelték a j-edik rendellenességet;} \\ 0, & \text{ha nem.} \end{cases}$$

$(j=1,\ldots,n;\ i=1,\ldots,m)$.

A gyerekek rendellenességeinek száma $\ell=1,2,\ldots,7$ lehetett, a pontosan ℓ számu rendellenességgel rendelkező gyerekeket pedig "ℓ-es csakok"-nak neveztük. Ennek alapján a B mátrix a következő:

$$b_{jj}=s_j - \sum_{\ell=1}^{m} \frac{a_{j\ell}^2}{t_\ell}$$

$$= (\text{a j-edik rendellenességben szenvedő és legalább még egy rendellenességgel rendelkező gyerekek száma})$$

$$- \frac{(\text{a j-edik rendellenességben szenvedő "2-es csakok" száma})}{2}$$

$$\vdots$$

$$- \frac{(\text{a j-edik rendellenességben szenvedő "7-es csakok" száma})}{7}$$

$$j=1,\ldots,n\ ;$$

$$b_{ji}=-\sum_{\ell=1}^{m} \frac{a_{j\ell}a_{i\ell}}{t_\ell}$$

$$= -\frac{(\text{az (i,j) kombinációban szenvedő "2-es csakok" száma})}{2}$$

$$\vdots$$

$$\frac{\text{(az i,j kombinációkban szenvedő "7-es csakok" száma)}}{7}$$

$$(1 \leq j < i \leq n),$$

és B szimmetrikus. Látható tehát, hogy elég a legalább két rendellenességgel rendelkező gyerekeket tekinteni, így $t_i > 1$, $(i=1,\ldots,m)$. Mivel minden rendellenesség előfordul valamely gyereken, $s_j > 0$ is igaz $(j=1,\ldots,n)$. (1) minimalizálásához az

$$XSX^T = \sum_{j=1}^{n} s_j \underline{x}_j \underline{x}_j^T = I_k$$

mellékfeltételt használtuk, ami a rendellenességek sulyozását jelenti a megfelelő rendellenességgel rendelkező gyerekek számával.

Mivel a rendellenességeket a síkban akartuk clusterezni, k=3 értéket választottunk. Az ábra azonban nem volt eléggé szemléletes: néhány rendellenesség-társulást sikerült ugyan elkülöniteni, a rendellenességek többsége azonban egy csomóban maradt. Ennek a rendellenességek heterogenitása lehet az oka, azaz hogy különböző gyerekmintákon különböző módon jelentkeznek és társulnak. Ez adta az ötletet az un. több lemezes beágyazáshoz:
Legyen $1<\upsilon<m$ természetes szám, és osszuk a gyerekeket ennyi diszjunkt csoportra. /Esetünkben célszerü υ-t 3-10 között választani, ahány diszjunkt csoportot el tudunk különiteni./ Ezzel azt mondjuk, hogy a gyerekeket υ lemezre rakjuk szét, de a 40 rendellenesség továbbra is szerepel mindegyik lemezen. Ha az ℓ-edik lemezen levő gyerekek száma m_ℓ, akkor $A^{(\ell)}$ nxm_ℓ-es indicencia-mátrix. Ezzel a beágyazást minden lemezen végrehajtva a rendellenességek és a gyerekek képeire az ℓ-edik lemezen az $\underline{x}_1^{(\ell)}, \ldots, \underline{x}_n^{(\ell)}$ ill. $y_1^{(\ell)}, \ldots, y_{m_\ell}^{(\ell)}$ pontokat kapjuk, melyek ténylegesen a síkban helyezkednek el.
Ezután a gyerekeket ismét szétosztjuk a lemezek között, mégpedig ugy, ha az i-edik gyereknek a $j_1,\ldots,j_p$ rendellenességei vannak, akkor arra a lemezre helyezzük, amelyre

$$\left\| \underline{Y}_i - \frac{\underline{x}_{j_1}^{(\ell)} + \ldots + \underline{x}_{j_p}^{(\ell)}}{p} \right\|, \quad (\ell=1,\ldots,\upsilon) \qquad (6)$$

minimális /ℓ-ben/.

Ezzel a felosztással az előbbieket szóról szóra megismételjük. Nyilvánvaló, hogy minden gyerekre a (6) kifejezés ℓ-ben való minimuma az egyes iterációs lépések során monoton csökken.
Az iteráció addig tart, amig ez az érték minden gyerekre egy ésszerüen választott K küszöb alá nem csökken. A többlemezes beágyazásra vonatkozó programfuttatások még folyamatban vannak.